Lecture Speech Recognition by Combining Word Graphs of Various Acoustic Models

Tetsuo Kosaka, Keisuke Goto, Takashi Ito and Masaharu Kato

Graduate School of Science and Engineering, Yamagata University,
Yonezawa-city, Yamagata, Japan

tkosaka@yz.yamagata-u.ac.jp

Abstract

The aim of this work is to improve the performance of lecture speech recognition by using a system combination approach. In this paper, we propose a new combination technique in which various types of acoustic models are combined. In the combination approach, the use of complementary information is important. In order to prepare acoustic models that incorporate a variety of acoustic features, we employ both continuous-mixture hidden Markov models (CMHMMs) and discrete-mixture hidden Markov models (DMHMMs). These models have different patterns of recognition errors. In addition, we propose a new maximum mutual information (MMI) estimation of the DMHMM parameters. In order to evaluate the performance of the proposed method, we conduct recognition experiments on "Corpus of Spontaneous Japanese." In the experiments, a combination of CMHMMs and DMHMMs whose parameters were estimated by using the MMI criterion exhibited the best recognition performance.

Index Terms: speech recognition, system combination, LVCSR, discrete-mixture HMM, word graph

1. Introduction

Although many studies have focused on improving the performance of large vocabulary continuous speech recognition (LVCSR) technologies, it is well known that the performance of these technologies is rather poor. The aim of this work is to improve the performance of lecture speech recognition by using a system combination approach. The basic concept of this approach is to use complementary information from several systems by combining their recognition results. Several types of combination techniques, such as recognizer output voting error reduction (ROVER) [1] and confusion network combination (CNC) [2] have been proposed. These combination techniques are often used for integrating word lattices from several different sites [1][3]. Examples of intra-site combinations are feature combination techniques where multiple acoustic features are combined [4][5]. Some researchers have attempted to carry out acoustic model combinations. In [6], acoustic models are trained in a different way to build systems that are complementary to each other. Evermann et al. have proposed a new combination approach in which triphone and quinphone HMMs trained by different training criteria are combined[7].

In this paper, we propose a new combination technique in which various types of acoustic models including not only continuous-mixture hidden Markov models (CMHMMs) but also discrete-mixture hidden Markov models (DMHMMs) are combined. In the combination approach, the use of complementary information is important. Therefore, a combination of various acoustic models that have different patterns of recognition errors is expected to improve the recognition performance. In order to build acoustic models that have a variety of acoustic features, we employ both CMHMMs and DMHMMs. Most speech recognition systems use CMHMMs as acoustic models because a conventional discrete HMM system based on vector quantization has a problem in that it is affected quantization distortion. In contrast, the DMHMM system is less affected by such distortion because the quantization size can be reduced by using a subvector or scalar quantization [8][9]. In our previous work [10], we demonstrated that the combination of CMHMMs and DMHMMs by ROVER is effective for noisy speech recognition in a medium size vocabulary task (5k words). In this paper, we build several acoustic models derived from various estimation criteria such as maximum likelihood (ML) estimation, maximum *a posteriori* (MAP) estimation, and maximum mutual information (MMI) estimation. In order to improve the performance of the lecture speech recognition task, we combine these models using the word graph combination technique [5]. A MAP estimation for the DMHMM parameters has already been proposed in [11]. In this paper, we propose an MMI estimation technique for DMHMM. In addition, we study objective indicators that are capable of objectively finding an effective combination of acoustic models, and we demonstrate that the phoneme mismatch rate (PMR) is useful as the objective indicator.

In order to evaluate the performance of the proposed method, we conduct recognition experiments on a spontaneous speech database "Corpus of Spontaneous Japanese" (CSJ). This corpus is the largest speech corpus in Japan and consists of approximately 7M words with a total speech length of 650 h [12].

2. Maximum mutual information estimation of DMHMM parameters

2.1. Discrete-mixture HMM

A DMHMM system requires a smaller amount of training data than a conventional DHMM system because the quantization size can be reduced by using a subvector or scalar quantization [8][9]. In the subvector-based method, feature vectors are partitioned into subvectors and are quantized using separate codebooks. In this study, we have used subvector quantization.

In the subvector quantization-based DMHMM, a feature vector $\boldsymbol{o}_t$ is partitioned into S subvectors, $\boldsymbol{o}_t = [\boldsymbol{o}_{1t}, \ldots, \boldsymbol{o}_{st}, \ldots, \boldsymbol{o}_{St}]$. VQ codebooks are provided for each subvector, and then the feature vector $\boldsymbol{o}_t$ is quantized,

$$q(\boldsymbol{o}_t) = [q_1(\boldsymbol{o}_{1t}), \ldots, q_s(\boldsymbol{o}_{st}), \ldots, q_S(\boldsymbol{o}_{St})]. \quad (1)$$

The output distribution of the DMHMM, $b_i(\boldsymbol{o_t})$, is given by

$$b_i(\boldsymbol{o}_t) = \sum_m w_{im} \prod_s \theta_{sim}(q_s(\boldsymbol{o}_{st})), \tag{2}$$

where w_{im} is the mixture coefficient for the m-th mixture in state i, and θ_{sim} is the probability of the discrete symbol for the s-th subvector.

2.2. Maximum likelihood estimation of DMHMM

The maximum likelihood (ML) estimate of the discrete probability $\theta^{ML}_{sim}(k)$ is calculated as follows:

$$\theta^{ML}_{sim}(k) = \frac{\gamma_{simk}}{\sum_{k'} \gamma_{simk'}}, \tag{3}$$

where γ_{simk} is given by

$$\gamma_{simk} = \sum_{t=1}^{T} \gamma_{im}(t) \delta(q_s(\mathbf{o}_{st}), k) \tag{4}$$

$$\delta(q_s(\mathbf{o}_{st}), k) = \begin{cases} 1 & q_s(\mathbf{o}_{st}) = k \\ 0 & otherwise \end{cases}.$$

k is the index of the subvector codebook and $\gamma_{im}(t)$ is the EM count of the m-th mixture component that is in state i at time t.

2.3. MAP estimation of DMHMM

In the ML estimation, the effect of the prior distribution is ignored. In contrast, an appropriate prior distribution is used in the MAP estimation [11]. The MAP estimate is given by

$$\theta^{MAP}_{sim}(k) = \frac{\tau_M \cdot \theta_{sim}(k) + n_{im} \cdot \theta^{ML}_{sim}(k)}{\tau_M + n_{im}} \tag{5}$$

$$n_{im} = \sum_{k'} \gamma_{simk'}, \tag{6}$$

where $\theta_{sim}(k)$ is the constrained prior parameter and τ_M indicates the relative balance between the corresponding prior parameter and the observed data. In our experiments, τ_M was set to 50.0.

2.4. MMI estimation of DMHMM

In this paper, we propose a technique for the MMI estimation of the DMHMM parameters. An MMI estimation for conventional discrete HMM (DHMM) parameters has been proposed in [13]. The differences between conventional DHMM and DMHMM are mixture representation and subvector partition. The re-estimation formula for objective functions for DMHMMs is as follows:

$$\theta^{MMI}_{sim}(k) = \frac{\theta_{sim}(k)(\frac{\partial \mathcal{F}}{\partial \theta_{sim}(k)} + E)}{\sum_{k'} \theta_{sim}(k')(\frac{\partial \mathcal{F}}{\partial \theta_{sim}(k')} + E)}. \tag{7}$$

The partial derivative can be given as

$$\frac{\partial \mathcal{F}}{\partial \theta_{sim}(k)} = \frac{1}{\theta_{sim}(k)}(\gamma_{simk} - \gamma^{gen}_{simk}), \tag{8}$$

where γ_{simk} is the standard EM count and γ^{gen}_{simk} represents the corresponding EM count obtained using the general model. E is a constant used for controlling the convergence speed. The division by small $\theta_{sim}(k)$ often causes the corresponding gradient coordinate to have a large magnitude in Eq. (8). In order to avoid this problem, we use the following approximation proposed by Merialdo [14]:

$$\frac{\partial \mathcal{F}}{\partial \theta_{sim}(k)} \approx \frac{\gamma_{simk}}{\sum_{k'} \gamma_{simk'}} - \frac{\gamma^{gen}_{simk}}{\sum_{k'} \gamma^{gen}_{simk'}} \tag{9}$$

By setting a smaller value of E in Eq. (7), we can increase the convergence speed; however, there is no theoretically proven convergence. Further, the training may get unstable. Therefore, we set the value by using the method proposed by Gopalakrishnan *et al.* [15]:

$$E = \max_{\theta_{sim}(k)} \left\{ -\frac{\partial \mathcal{F}}{\partial \theta_{sim}(k)}, 0 \right\} + \epsilon_e, \tag{10}$$

where ϵ_e is a small positive constant. Mixture weights can be calculated by the MMI estimation in a similar way. Phoneme graphs generated for all training utterances are used as a compact representation of many different hypotheses. The EM counts of the general models are calculated using these hypotheses.

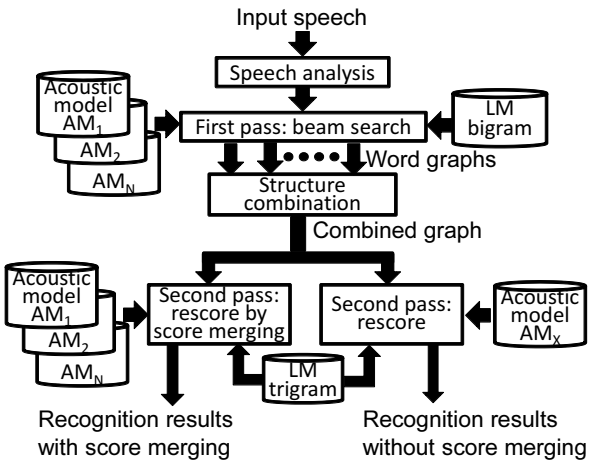

Figure 1: *Block diagram of word graph combination.*

3. Word graph combination using DMHMMs and CMHMMs

In the proposed method, we use a word graph combination approach to integrate the intermediate outputs of several types of acoustic models. It is well known that a system combination approach has shown significant improvements over the results with just a single system if the recognition results are sufficiently different among the systems. In order to improve the recognition performance, the outputs of DMHMMs and CMHMMs are integrated. Some system combination approaches (e.g., ROVER and CNC) are proposed to improve the performance of LVCSR systems. In this study, we use the word graph combination technique that has been proposed by Chen et al. [5]. In this combination technique, word graphs are directly integrated. Unlike in conventional combination approaches such as ROVER or CNC, in this technique, the timing information for all word hypotheses is well preserved.

Fig. 1 shows the block diagram of the proposed system. A one-pass frame-synchronous search algorithm using beam searching has been adopted in the first pass. The search algorithm calculates the acoustic and language likelihood of obtaining word graphs. In the first pass, several acoustic models (e.g. DMHMM and CMHMM) are used, and a bigram is used as a language model. After several word graphs are obtained;

they are combined to form a single graph. In the second pass, two rescoring methods are compared in the experiments. One method employs several acoustic models, and the other employs a single acoustic model for rescoring. In the former method, several scores derived from the different acoustic models are merged. The merged score is simply obtained by averaging all scores. In the latter method, the combined graph is rescored by the single acoustic model. In this pass, a trigram is also used as the language model.

Here, we describe the algorithm of the word graph combination. Suppose that there are N word graphs, $W_1, W_2, \ldots, W_N$, to be combined. If two arcs q_1 in W_1 and q_2 in W_2 are equal, the two word graphs W_1 and W_2 can be combined as

$$\begin{aligned} & W_1 + W_2 \\ = \; & \{q = q_1 + q_2 | q_1 = q_2\} \cup \{q_1 | q_1 \notin W_2\} \\ & \cup \{q_2 | q_2 \notin W_1\}. \end{aligned} \tag{11}$$

Two equal arcs have the same word ID, start time, and end time. The word graph combination for N systems can be obtained as

$$W = W_1 + W_2 + \cdots + W_N = \sum_{i=1}^{N} W_i. \tag{12}$$

4. Experimental set-up

In this section, we describe the LVSCR system, which is used for recognition experiments. In the speech analysis module, a speech signal is digitized at a sampling frequency of 16 kHz with a quantization size of 16 bits. The length of the analysis frame is 25 ms and the frame period is set to 8 ms. A 13-dimensional feature (12-dimensional MFCC and log power) is derived from the digitized samples for each frame. Further, the delta and the delta-delta features are calculated from the MFCC feature and the log power. Then, the total number of dimensions is 39. The 39-dimensional parameters are normalized by using the cepstral mean normalization (CMN) method. A two-pass search decoder using a bigram and a trigram is used for recognition. Decoding is performed using a one-pass algorithm in which a frame-synchronous beam search and a tree-structured lexicon are applied in the first pass. The bigram and trigram models are trained from text data containing 2,668 lectures in the CSJ, and the total number of words is 6.68M. Trained language models have 47,099 word-pronunciation entries.

The total number of lectures used for acoustic model training is 963, and the total speech length is 203 h. One lecture is given by one speaker; therefore, the total number of speakers is 963. Note that some speakers give several lectures. Five types of acoustic model sets are used; these models are built for the word graph combination (see Table 1). A set of shared-state triphones is used as the acoustic model for each model scheme. The number of states and the number of mixture components are the same for each model set. The number of states is 3000, and the number of mixture components is 16. Table 2 shows the subvector allocation and the codebook size for DMHMMs. In the table, although Δ and Δ^2 are omitted, these codebooks are designed in the same manner.

We use the "testset1" evaluation set, which consists of academic presentations given by 10 male speakers. This is one of the standard test sets in the CSJ. The experimental results of each research group can be compared by using this test set. The total speech length is 1.7 h.

Table 1: *Conditions of acoustic models.*

Identifier	Model	Estimation
CMHMM-ML	CMHMM	ML
CMHMM-MMI		MMI
DMHMM-ML	DMHMM	ML
DMHMM-MAP		MAP
DMHMM-MMI		MMI

Table 2: *Codebook size for each subvector.*

Parameter	logP	c_1, c_2	c_3, c_4	c_5, c_6	c_7, c_8	c_9, c_{10}	c_{11}, c_{12}
CB size	64	64	64	64	64	64	64

5. Results and discussions

The recognition results of the proposed combination technique using CMHMM-MMI and DMHMM-MMI are shown in Fig. 2. The word error rates (WERs) of CMHMM-MMI and DMHMM-MMI without combination are 20.78% and 20.57%, respectively. The performance of DMHMM-MMI is slightly better than that of CHMM-MMI; however, the difference is small. The results of word combination without score merging are shown in this figure. In this method, the word graph of CMHMM-MMI and that of DMHMM-MMI are combined after the first pass processing. In the second pass, scores are given by using either CMHMM-MMI or DMHMM-MMI. The WER of this method is 20.31% for CMHMM-MMI and 20.53% for DMHMM-MMI. These results show a better performance than the results without combination. The method of word combination with score merging obtained the best performance (19.88%).

Various combinations of the two models have been tested, and the results are shown in Table 3. The performance of all the methods except the combination of DMHMM-MMI and DMHMM-MAP can be improved. The combination of CMHMM-MMI and DMHMM-MMI exhibits the best performance (19.88%). In contrast, the combination of DMHMM-MMI and DMHMM-MAP exhibits the worst performance. What is interesting is that the average WER of CMHMM-MMI (20.78%) and DMHMM-MMI (20.57%) before the combination is the same as that of DMHMM-MMI (20.57%) and DMHMM-MAP (20.78%). However, the performance after

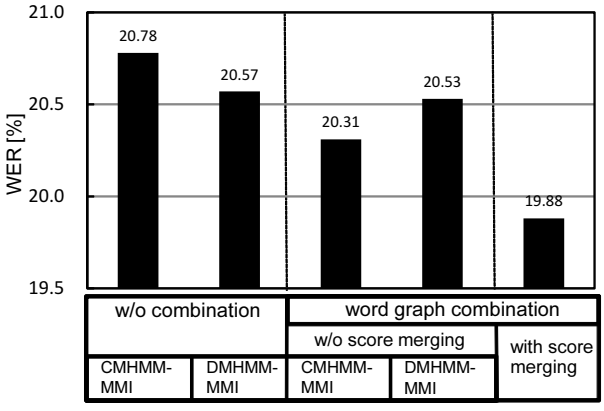

Figure 2: *Recognition results of word graph combination using CMHMM-MMI and DMHMM-MMI.*

Table 3: *Results of word graph combination using various acoustic models [%].*

Model 1	WER	Model 2	WER	Average WER before combination	WER after combination	PMR	Improvement
CMHMM-ML	21.33	CMHMM-MMI	20.78	21.05	20.20	7.02	4.04
		DMHMM-ML	20.61	20.97	20.47	4.74	2.38
		DMHMM-MAP	20.78	21.05	20.57	4.62	2.28
CMHMM-MMI	20.78	DMHMM-ML	20.61	20.70	19.89	7.55	3.91
		DMHMM-MMI	20.57	20.67	**19.88**	7.17	3.82
		DMHMM-MAP	20.78	20.78	19.98	7.42	3.85
DMHMM-MMI	20.57	DMHMM-MAP	20.78	20.67	20.64	0.97	0.15

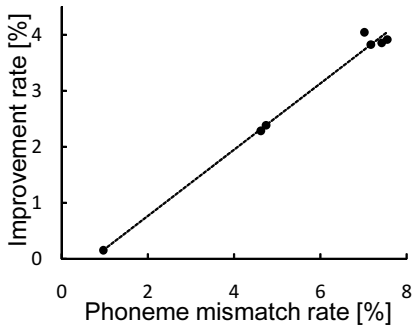

Figure 3: *Relation between phoneme mismatch rate and improvement rate.*

combination of the latter is worse than that of the former. This is because the pattern of recognition error is similar between DMHMM-MMI and DMHMM-MAP. This suggests that a combination of models that have different patterns of recognition errors is important for improving the performance.

In order to clarify this point, we have calculated a phoneme mismatch rate (PMR) between two models. For the PMR calculation, the two phoneme recognition results are aligned, and the differences in the phoneme level are counted. If the PMR is equal to zero, the two results are the same. The relation between PMR and the improvement rate is plotted in Fig. 3. The improvement rate is calculated using the average WER before the combination and the WER after the combination. From this figure, we can see the improvement rate and PMR show a strong correlation. The results suggest that the difference in error patterns is important to improve the recognition performance.

Further, on the basis of the tests of a combination of three or more systems carried out by using the acoustic models shown in Table 1, we conclude that a better performance cannot be obtained.

6. Conclusions

This paper proposed a new combination technique in which various types of acoustic models including not only CMHMMs but also DMHMMs are combined. This is because these models have different patterns of recognition errors. For the system combination, we employed the word graph combination technique. From the recognition experiments, we observed that the combination of CMHMM-MMI and DMHMM-MMI exhibited the best performance among several combination methods. In addition, we demonstrated that the phoneme mismatch rate (PMR) was useful information for finding an effective combination of acoustic models. We plan to compare our approach based on word graph combination with other combination techniques such as ROVER or CNC.

7. References

[1] J. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER)," in *Proc. of the IEEE Automatic Speech Recognition and Understanding Workshop*, 1997, pp. 347 – 352.

[2] L. Mangu, E. Brill, and A. Stolcke, "Finding consensus in speech recognition: Word error minimization and other applications of confusion networks," *Computer Speech and Language*, vol. 14, pp. 373 – 400, 2000.

[3] B. Hoffimeister, R. Schluter, and H. Ney, "iCNC and iROVER: The limits of improving system combination with classification?," in *Proc. of Interspeech2008*, 2008, pp. 232–235.

[4] R. Schlüter, I. Bezrukov, H. Wagner, and H. Ney, "Gammatone features and feature combination for large vocabulary speech recognition," in *Proc. ICASSP2007*, 2007, pp. 649–652.

[5] I.-F. Chen and L.-S. Lee, "A new framework for system combination based on integrated hypothesis space," in *Proc. of Interspeech2006*, 2006, pp. 533 – 536.

[6] C. Breslin and M.J.F. Gales, "Generating complementary systems for speech recognition," in *Proc. of Interspeech2006*, 2006, pp. 525–528.

[7] G. Evermann and P. Woodland, "Posterior probability decoding, confidence estimation and system combination," in *NIST Speech Transcription Workshop*, 2000.

[8] S. Tsakalidis, V. Digalakis, and L. Newmeyer, "Efficient speech recognition using subvector quantization and discrete-mixture HMMs," in *Proc. of ICASSP99*, 1999, pp. 569–572.

[9] S. Takahashi, K. Aikawa, and S. Sagayama, "Discrete mixture HMM," in *Proc. of ICASSP97*, 1997, pp. 971–974.

[10] T. Kosaka, Y. Saito, and M. Kato, "Noisy speech recognition by using output combination of discrete-mixture HMMs and continuous-mixture HMMs," in *Proc. of Interspeech2009*, 2009, pp. 2379–2382.

[11] T. Kosaka, M. Katoh, and M. Kohda, "Noisy speech recognition with discrete-mixture HMMs based on MAP estimation," in *Proc. of ICA2004*, 2004, vol. 2, pp. 1691–1694.

[12] S.Furui, M.Nakamura, T.Ichiba, and K.Iwano, "Analysis and recognition of spontaneous speech using corpus of spontaneous Japanese," *Speech Communication*, vol. 47, pp. 208 – 219, 2005.

[13] L. R. Bahl, P. F. Brown, P. V. de Souza, and R. L. Mercer, "Maximum mutual information estimation of hidden Markov model parameters for speech recognition," in *Proc. of ICASSP86*, 1986, pp. 49–52.

[14] B. Merialdo, "Phonetic recognition using hidden Markov models and maximum mutual information training," in *Proc. of ICASSP88*, 1988, pp. 111–114.

[15] P. S. Gopalakrishnan, D. Kanevsky, A. Nadas, and D. Nahamoo, "A generalization of the Baum algorithm to rational objective functions," in *Proc. of ICASSP89*, 1989, pp. 631–634.